

LA-UR-21-27447

Approved for public release; distribution is unlimited.

Title: An Exploration of Ensemble GI software using LANL TA-66 sensor data

Author(s): Woodring, Jonathan Lee

Intended for: NA-221 Objective L Data Analytics Project Team Status Update

Issued: 2021-08-02 (rev.1)

Disclaimer:

Los Alamos National Laboratory, an affirmative action/equal opportunity employer, is operated by Triad National Security, LLC for the National Nuclear Security Administration of U.S. Department of Energy under contract 89233218CNA000001. By approving this article, the publisher recognizes that the U.S. Government retains nonexclusive, royalty-free license to publish or reproduce the published form of this contribution, or to allow others to do so, for U.S. Government purposes. Los Alamos National Laboratory requests that the publisher identify this article as work performed under the auspices of the U.S. Department of Energy. Los Alamos National Laboratory strongly supports academic freedom and a researcher's right to publish; as an institution, however, the Laboratory does not endorse the viewpoint of a publication or guarantee its technical correctness.

An Exploration of Ensemble GI software using LANL TA- 66 sensor data

Jon Woodring, LANL

July 29, 2021

LA-UR-21-27447

Outline

- Process of testing
- Smoothing Data
- Data gap (missing data) experiment, no resampling for sparsities
- Data gap (missing data) experiment, resampled data for dense time
- Conclusion
- **Note about graph labels in the following slides:**
 - **x axis is always time**
 - **y axis is sensor value, *except* on discrete quantile bin plots, which are ordinal, i.e., the nth varying width quantile bin, not percentages**

Process of testing

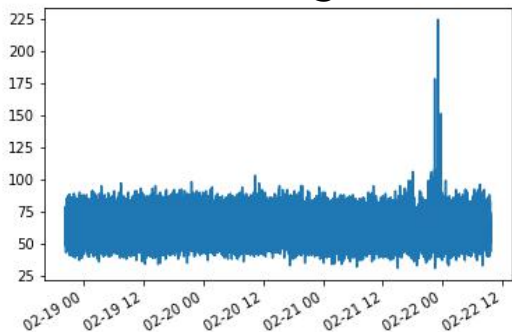
- Acquired Ensemble Grammar Induction (GI) software
- Tested with raw data using two years of time series
- Consulted with Constantin Brif
- Smoothed data with Lasso
- Tested 3 Lasso'd sensor data sets in Ensemble GI GUI
- Data gap (missing data) experiment
 - TA-66 sensor data has missing values
 - Find data gaps as anomalies? sparse time
 - Find data gaps as anomalies? dense time (reinterpolating) the data in time

Smoothing data, i.e., reducing the data

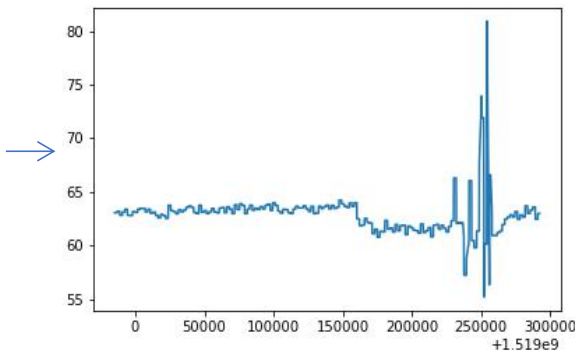
- Lasso (Least Absolute Shrinkage and Selection Operator) + Lars (Least Angle Regression)
 - Piecewise linear regression; used before on similar types of MINOS sensor data
 - Lines can be sloped, depending on Lasso parameter tuning
- Manual symbol (label) generation from Lasso line segments
 - Run Lasso to generate at most 1 line segment per 1 minute: there may be fewer segments where one line spans multiple minutes, depending on goodness of fit
 - Generate one feature per minute as the average of the line over that minute
 - Preselect a number of symbols (**n**) and discretize each minute into **n** varying width quantile bins to make sensor values uniform (e.g., k bin discretizer)
- Ensemble GI settings to match
 - Piecewise aggregate approximation (PAA) = 1, no need to further linearize
 - Number of symbols = **n**, no need to further discretize

1 minute Lasso segments and 20 bins over several days

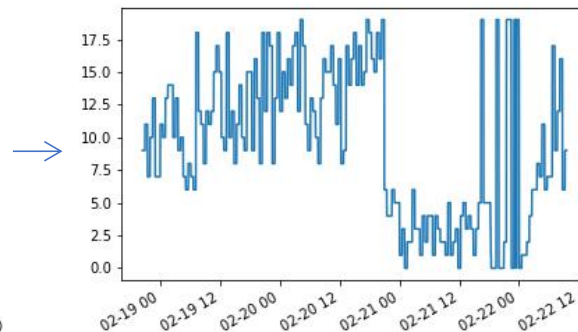
Neutron Singles Raw



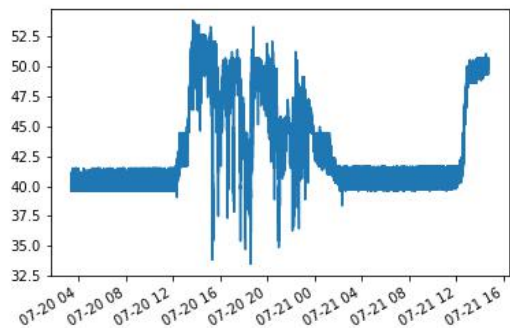
Lasso



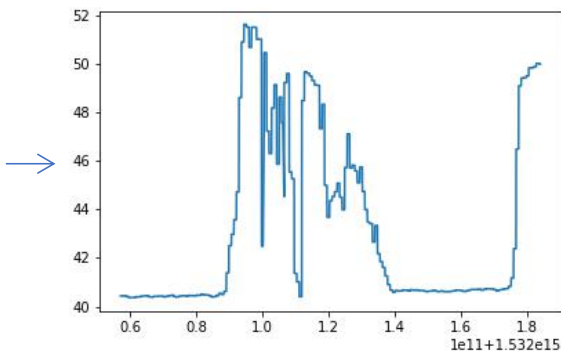
Labeled Quantile Bins



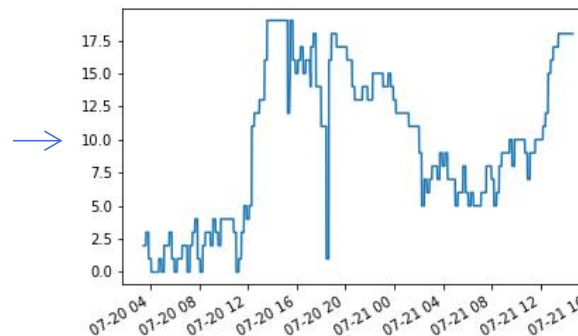
Light 3 Raw



Lasso



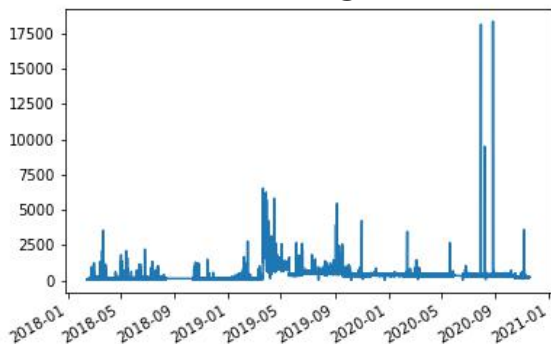
Labeled Quantile Bins



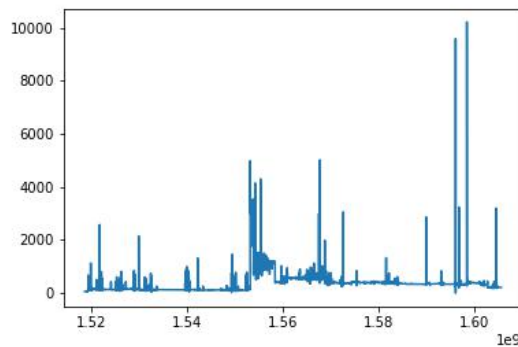
Data Set 1

- Neutron detector
 - 25 million rows, two columns
 - 2 years at ~3 second intervals, single and double counts ; example **gaps** in data
 - Smoothed to 1 minute intervals (20:1 in time), 20 discrete quantiles

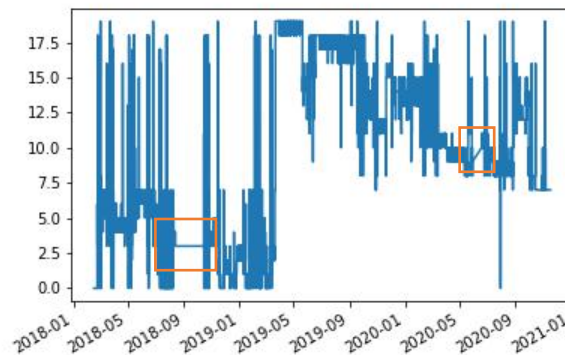
raw singles



Lasso + Lars



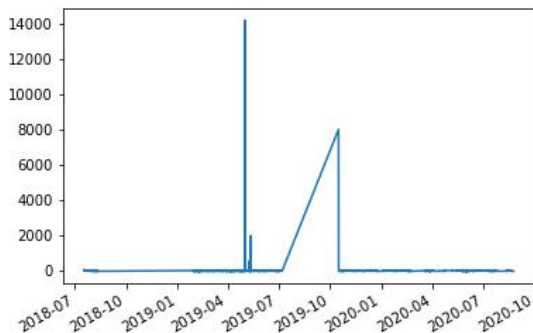
quantile bins



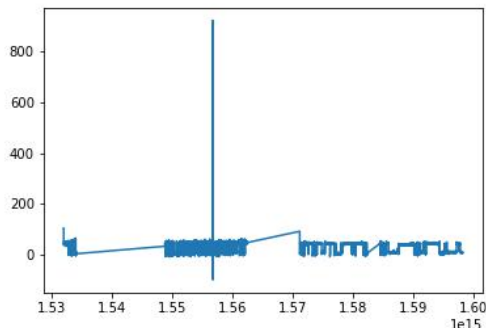
Data Set 2

- Light Sensor 3 (light closest to neutron detector)
 - Raw 31 million rows, one column
 - 2 years of ~1 second intervals, light intensity ; example **gap(s)** in data
 - Smoothed to 1 minute intervals (60:1 in time), 20 discrete quantiles

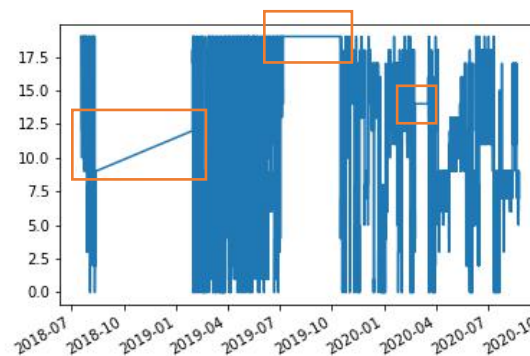
raw singles



Lasso + Lars



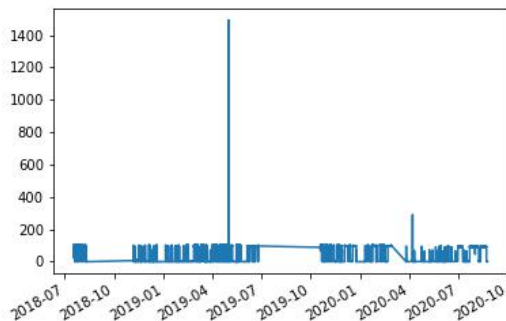
quantile bins



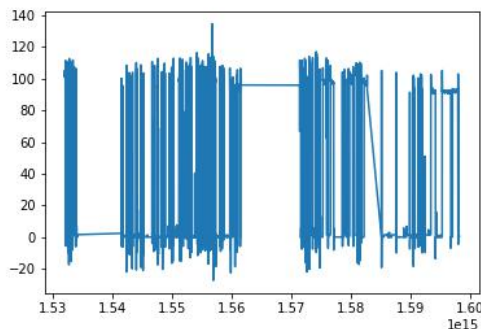
Data Set 3

- Light Sensor 6 (1 of 3 conference room sensors)
 - 35 million rows, one column
 - 2 years of 1 second intervals, light intensity ; example **gap**(s) in data
 - Smoothed to 1 minute intervals (60:1 in time), 9 discrete quantiles

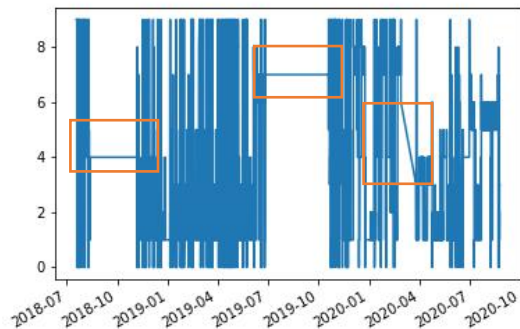
raw singles



Lasso + Lars



quantile bins

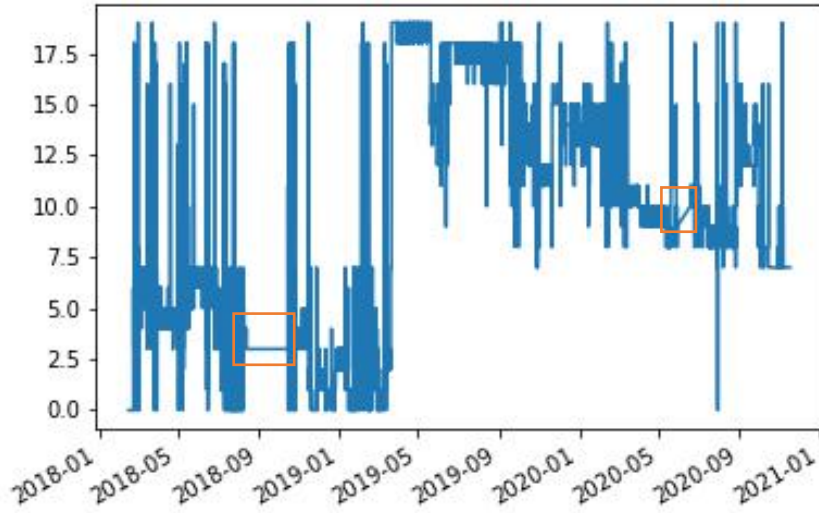


Experiment to find missing data, irregular time

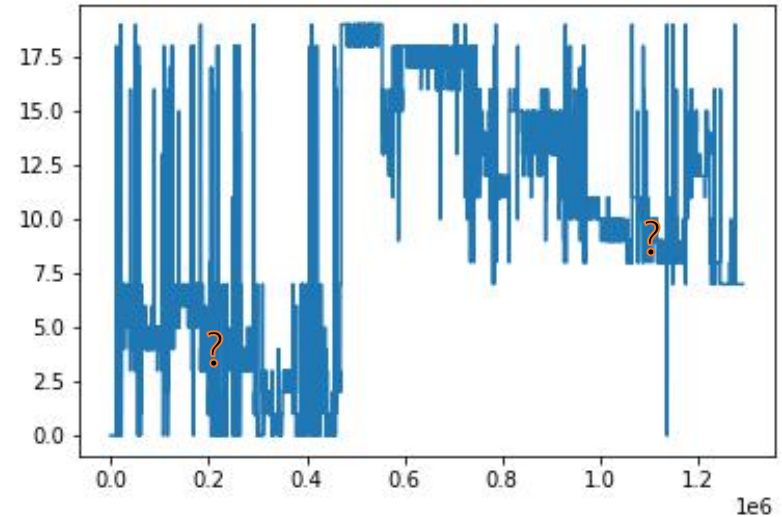
- All of the TA-66 data streams do not have regularly sampled data
 - **Data are sparse, in general, due to possible lag in sensor timing or events**
 - Also, power outage, sensor/platform failure, data collection interrupted, etc.
 - Our data are time stamped for each sample: i.e., **you cannot infer real time from sample position in the stream due to irregular sampling**
- Curious to see if the anomaly detector could flag missing data
 - Tested sensor data **as is** with Ensemble GI: e.g., the streams were unevenly sampled in time, due to drops or just sensor time lag between samples
- **Experiment 1: Would missing data be flagged without real time?**
 - **Would discontinuities in the value space (sensor readings) “a data drop blip” would be noticable without the time index/being densely, regularly sampled?**

“Compressed in time”: i.e., sparse real time vs. indexed time

sparse but timestamped



sparse, no timestamp

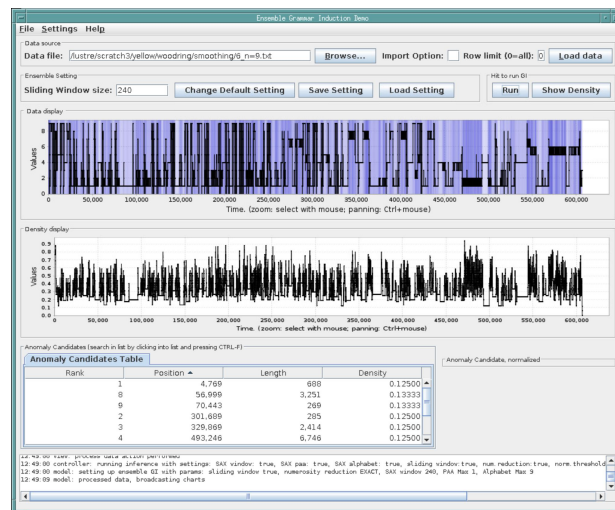
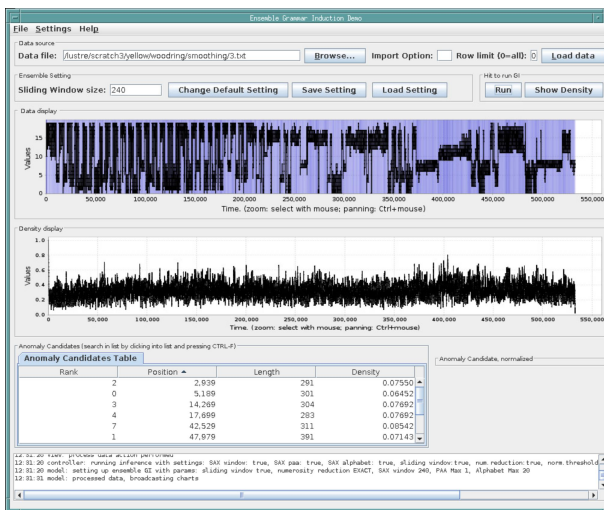


Running Ensemble GI with sparse time data

neutron singles

light sensor 3

light sensor 6

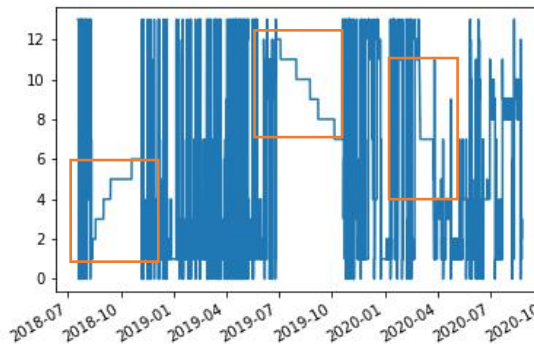
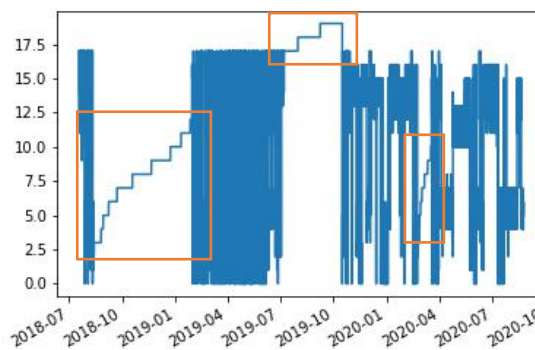
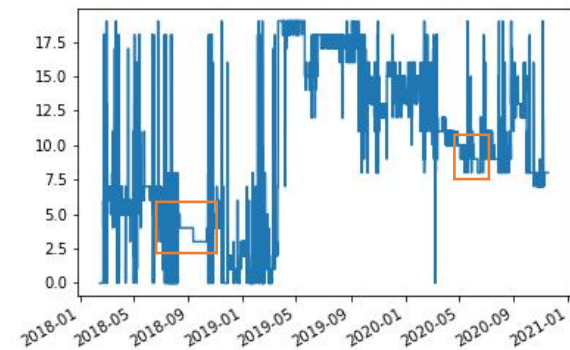
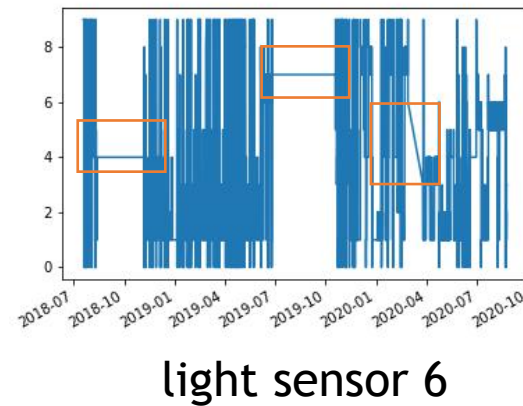
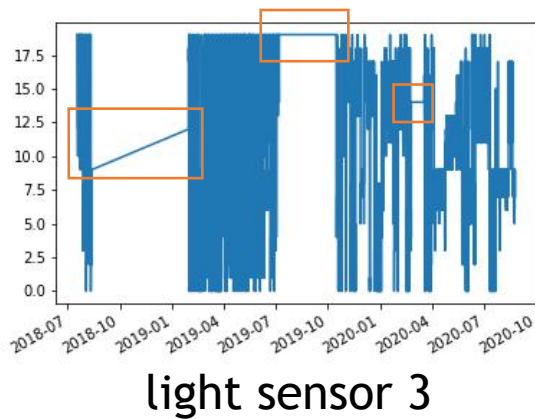
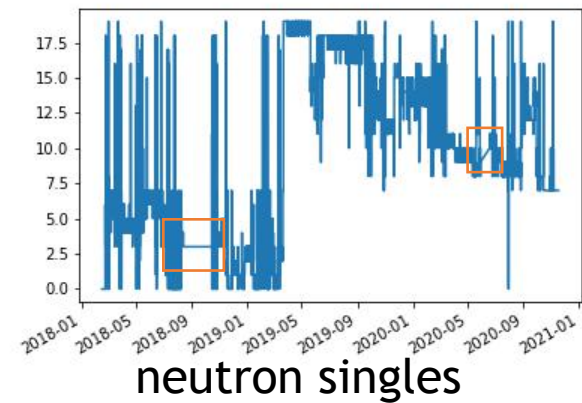


- 3 trials, 10 top anomalies, windows of 2 minutes, 15 minutes, 4 hours, 1 day
- Didn't seem to find anomalies to that matched time of gaps (**User error?**)

Rerun experiment with dense sampled in time data

- **User error? Should have just smoothed and downsampled without uniform binning? Unanswered questions**
- **Experiment 2: Rerun, but with regular/evenly spaced samples in time**
 - Same settings for Lasso and quantile binning
 - Same settings for Ensemble GI, including number of trials and top anomalies
 - Additional data for Ensemble GI to process, but no noticeable addition in time
- **Resample the data to have regular (even) sampling in time**
 - Neutron data, reinterpolated to have exactly 1 sample / 3 seconds
 - Light sensor 3 and light sensor 6, interpolated to have exactly 1 sample / 1 second
 - Filled in gaps with samples, but those samples were anomalous wrt other data

Resampled comparison - sparse top, dense bottom

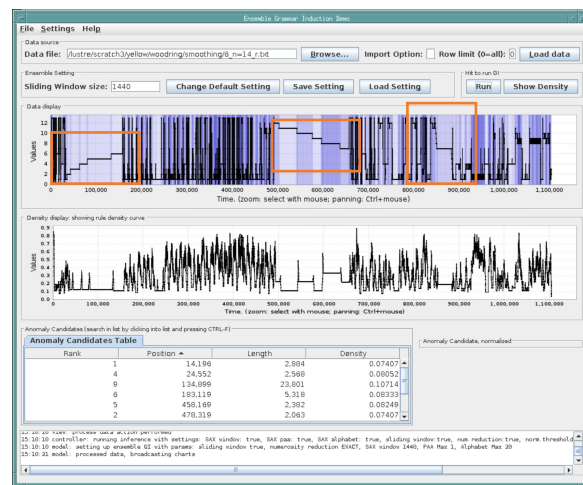
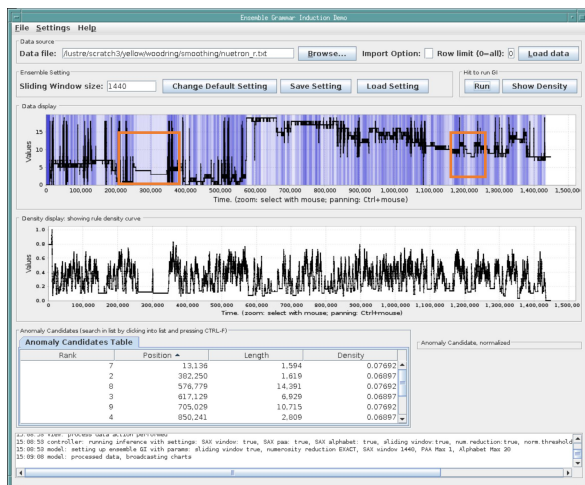


Running Ensemble GI with dense time (regular Hz)

neutron singles

light sensor 3

light sensor 6



- 3 trials, 10 top anomalies, windows of 2 minutes, 15 minutes, 4 hours, 1 day
- **With a time window of 1 day: Ensemble GI flagged the missing data, i.e., low rule density, i.e., anomalies, correspond to the missing data**

Example timings of missing data compared to anomalies

Missing data (from, to)

- Neutron detector
 - 2018-08-13 12:10:43 to
2018-10-12 21:35:16
- Light sensor 3
 - 2018-08-12 23:59:59 to
2019-01-29 06:12:56

Anomalies (initial, length in minutes)

- Neutron detector
 - 2018-08-14 12:11:00 41599
- Light sensor 3
 - 2018-08-14 01:07:55 16270

Conclusion

- Infrastructure in place to test with TA-66 data and first high-level results
- Fast anomaly detection with 2 years of data at 1 minute (< 30 seconds)
- Future Work
 - More testing
 - Integrate Ensemble GI to generate features
 - Build results into feature matrix
 - What to do about sparse data, in general?
 - Correlate anomalies across sensors and time scales